

# Joint Scheduling and ARQ for MU-MIMO Downlink in the Presence of Inter-Cell Interference

H.Shirani-Mehr\*, H. Papadopoulos<sup>†</sup>, S. A. Ramprashad<sup>†</sup>, G. Caire\*

\*University of Southern California, Email: shiranim, caire@usc.edu

<sup>†</sup>DoCoMo Laboratories USA, Inc., Email: ramprashad, hpapadopoulos@docomolabs-usa.com

## Abstract

User scheduling and multiuser multi-antenna (MU-MIMO) transmission are at the core of high-rate data-oriented downlink schemes of the next-generation of cellular systems (e.g., LTE-Advanced). Scheduling selects groups of users according to their channels vector directions and SINR levels. However, when scheduling is applied independently in each cell, the inter-cell interference (ICI) power at each user receiver is not known in advance since it changes at each new scheduling slot depending on the scheduling decisions of all interfering base stations. In order to cope with this uncertainty, we consider the joint operation of scheduling, MU-MIMO beamforming and Automatic Repeat reQuest (ARQ). We develop a game-theoretic framework for this problem and build on stochastic optimization techniques in order to find optimal scheduling and ARQ schemes. Particularizing our framework to the case of “outage service rates”, we obtain a scheme based on adaptive variable-rate coding at the physical layer, combined with ARQ at the Logical Link Control (ARQ-LLC). Then, we present a novel scheme based on incremental redundancy Hybrid ARQ (HARQ) that is able to achieve a throughput performance arbitrarily close to the “genie-aided service rates”, with no need for a genie that provides non-causally the ICI power levels. The novel HARQ scheme is both easier to implement and superior in performance with respect to the conventional combination of adaptive variable-rate coding and ARQ-LLC.

**Keywords:** Multiuser MIMO, inter-cell interference, scheduling, hybrid ARQ, stochastic optimization, game theory.

## I. INTRODUCTION

High-rate data-oriented downlink schemes [1], [2] have been successfully deployed as an extension of 3G cellular standards (WCDMA and CDMA2000). These schemes are based on the results of [3]–[5], showing that the throughput (or “ergodic”) sum-capacity of single-antenna multi-access (uplink) and broadcast (downlink) fading Gaussian channels is achieved by allocating opportunistically each time-frequency slot to the user with the best instantaneous channel conditions. In a multiuser setting, the sum-capacity is usually not the most meaningful measure of the system performance. Instead, maximizing the sum-throughput subject to some *fairness* constraint is more desirable [5]. To this purpose, a *downlink scheduling policy* can be designed in order to maximize a suitable concave and component-wise monotonically increasing network utility function over the system’s achievable throughput region (i.e., the region of achievable long-term average user rates). The network utility function is designed in order to capture the desired notion of “fairness” (e.g., proportional fairness, max-min fairness and, more in general,  $\alpha$ -fairness [6]).

In the next generation of cellular systems (e.g., the so-called LTE-Advanced [7]), high-rate data-oriented downlink schemes will be combined with multiuser multi-antenna (MU-MIMO) transmission techniques [8], [9], supporting spectral efficiencies in the 10’s of bits/sec/Hz [10], [11]. With MU-MIMO, the rate supported by each user is generally a function of *all* the user channel vectors, and depends on the type of MU-MIMO precoding [8], [9], [12]. In order to compute the transmitter precoder parameters (e.g., the beamforming steering vectors and the transmitted rates and powers), channel state information at the transmitter (CSIT) is required. This can be accurately obtained using open and closed loop channel estimation and feedback schemes (the literature on this subject is overwhelming, for example, see [13]–[17] and references therein).

In particular, scheduling with MU-MIMO and non-perfect CSIT was considered in [18], particularizing the general stochastic optimization framework of [19] to the case of a single-cell system with linear Zero-Forcing Beamforming (ZFBF) MU-MIMO precoding, where CSIT is obtained via noisy channel estimation and prediction.

In this work we focus on a multi-cell environment with no inter-cell cooperation. For sufficiently slowly-moving user terminals it is possible to design training and feedback schemes

that achieve almost perfect CSIT [16]–[18]. Therefore, for simplicity we shall assume that each BS has perfect CSIT for its own users. In contrast, in a multi-cell system, inter-cell interference (ICI) emerges as another source of unavoidable uncertainty. (see [20], [21] and references therein). When the schedulers at each BS make their own decisions independently, based only on the locally available CSIT relative to their own users, the ICI power seen at each user receiver changes on a slot-by slot basis in a random and unpredictable manner, depending on the scheduling decision made at all the interfering BSs. As a consequence, the instantaneous Signal to Interference plus Noise Ratio (SINR) “seen” at any given user receiver is a random variable.

The decentralized scheduling problem in a multi-cell environment can be formulated as a non-cooperative game: each BS (player) wishes to maximize its own utility function over its own feasible throughput region. The players’ strategies are all feasible scheduling policies. In addition, the throughput region of any given cell depends on the ICI power statistics seen at the users’ receivers, which in turn depend on the scheduling policies applied at the interfering BSs. We show that when the individual network utility functions are concave the multi-cell decentralized scheduling game is a concave game and therefore Nash equilibria exist.

In order to solve the network utility maximization at each BS, for given ICI statistics, we apply the stochastic optimization framework of [18], [19], [22], [23]. A straightforward application of this approach yields a scheme based on variable-rate adaptive coding at the physical layer, and conventional ARQ at the Logical Link Control (LLC) layer. We notice that similar approaches are included in several wireless standards such as EV-DO and HSDPA [24]–[26], and therefore this can be regarded as the base-line “conventional” approach. In order to improve upon the conventional approach, we propose a new method based on combining incremental redundancy Hybrid Automatic Retransmission reQuest (HARQ) [27] and MU-MIMO opportunistic scheduling. In the proposed scheme, each user feeds back the value of the instantaneous mutual information observed in the previous slot, that is used by the scheduler to update recursively the scheduler weights. We show that the throughput achieved by the proposed HARQ scheme approaches the throughput of a “virtual system”, as if a genie provided non-causally the ICI values at each scheduling slot. However, we stress that the proposed scheme makes use of strictly causal information, and therefore requires no genie.

## II. SYSTEM SETUP

We consider the downlink of a system with  $C > 1$  cells. In each cell, a BS equipped with  $M$  antennas transmits to  $K$  single-antenna users. The channel is assumed frequency flat<sup>1</sup> and constant over “slots” of length  $T \gg 1$  symbols (block-fading model [28]). Any given channel use of the complex baseband discrete-time signal at the  $k$ -th user in cell  $c$  during slot  $t$  is described by

$$y_{k,c}[t] = \underbrace{\sqrt{g_{k,c,c}} \mathbf{h}_{k,c,c}^H[t] \mathbf{x}_c[t]}_{\text{desired BS}} + \underbrace{\sum_{c' \neq c} \sqrt{g_{k,c,c'}} \mathbf{h}_{k,c,c'}^H[t] \mathbf{x}_{c'}[t]}_{\text{inter-cell interference}} + z_{k,c}[t], \quad (1)$$

where  $t$  ticks at the slot rate,  $(k, c)$  denotes user  $k$  in cell  $c$ ,  $\mathbf{h}_{k,c,c'}[t] \in \mathbb{C}^M$  is the channel vector from the  $c'$ -th BS antenna array to the  $(k, c)$ -th receiver antenna,  $\mathbf{x}_c[t] \in \mathbb{C}^M$  is the signal transmitted by  $c'$ -th BS and  $z_{k,c}[t] \sim \mathcal{CN}(0, 1)$  is the additive white Gaussian noise (AWGN) sample. The coefficients  $g_{k,c,c'}$  are distance-dependent path gains [29] that are assumed to be time-invariant over many slots. The BSs are sum-power constrained such that  $\text{tr}(\Sigma_c[t]) \leq 1$  for all  $t$ , where  $\Sigma_c[t] = \mathbb{E}[\mathbf{x}_c[t] \mathbf{x}_c^H[t]]$  denotes the transmit covariance matrix. The actual channel SNR is included as a common scaling factor in the coefficients  $g_{k,c,c'}$ . The channel vectors of users in cell  $c$  form the columns of the channel matrix  $\mathbf{H}_c[t] = [\mathbf{h}_{1,c,c}[t], \dots, \mathbf{h}_{K,c,c}[t]] \in \mathbb{C}^{M \times K}$ . We assume that all vectors  $\mathbf{h}_{k,c,c'}[t]$  are mutually independent with i.i.d. components  $\sim \mathcal{CN}(0, 1)$ , for all distinct 4-tuples  $(t, k, c, c')$ . Each BS  $c$  knows all time-invariant quantities relative to its own users and has perfect knowledge of  $\mathbf{H}_c[t]$  immediately before the beginning of slot  $t$  (perfect CSIT for the own users).

A feasible scheduling policy  $\gamma_c$  for BS  $c$  is a possibly *randomized* stationary function<sup>2</sup> that maps  $\mathbf{H}_c[t]$  into the pair  $\gamma_c(\mathbf{H}_c[t]) = (\Sigma_c[t], \mathbf{r}_c[t])$ , where  $\mathbf{r}_c[t] = (r_{1,c}[t], \dots, r_{K,c}[t])$  is a rate allocation vector. We assume that the MU-MIMO precoder is based on linear ZFBF. This yields the transmitted signal vector in the form  $\mathbf{x}_c[t] = \sum_{k \in \mathcal{S}_c[t]} \mathbf{v}_{k,c}[t] u_{k,c}[t]$ , where  $\mathcal{S}_c[t]$  denotes the set of active users, i.e., users that are selected to be served on slot  $t$  and where  $u_{k,c}[t] \in \mathbb{C}$  denotes the coded symbol for user  $(k, c)$ , with power  $\mathbb{E}[|u_{k,c}[t]|^2] = P_{k,c}[t]$ . The ZFBF steering vectors  $\{\mathbf{v}_{k,c}[t] : k \in \mathcal{S}_c[t]\}$  are given by the unit-norm (normalized)  $k$ -th column of the Moore-Penrose

<sup>1</sup>The generalization to MIMO-OFDM and frequency selective fading is immediate.

<sup>2</sup>Using the theory developed in [30] we can show that restricting to *stationary* policies does involve any suboptimality in terms of the achievable throughput region.

pseudoinverse (e.g., see [14], [16], [18], [31], [32] and references therein) of the channel matrix restricted to the active users, i.e., to the columns  $\{\mathbf{h}_{k,c,c}[t] : k \in \mathcal{S}_c[t]\}$ . It follows that the transmit covariance matrix takes on the form

$$\Sigma_c[t] = \sum_{k \in \mathcal{S}_c[t]} \mathbf{v}_{k,c}[t] \mathbf{v}_{k,c}^H[t] P_{k,c}[t]. \quad (2)$$

where non-negative coefficients  $\{P_{k,c}[t] : k \in \mathcal{S}_c[t]\}$  define the power allocation over the active users in cell  $c$ , and satisfy the power constraint  $\sum_{k \in \mathcal{S}_c[t]} P_{k,c}[t] \leq 1$ . A necessary and sufficient condition for perfect zero-forcing of the intra-cell multiuser interference is that  $|\mathcal{S}_c[t]| \leq \min\{M, K\}$ . Without loss of generality, in the following we identify the set of active users  $\mathcal{S}_c[t]$  with those users with positive powers, i.e.,  $P_{k,c}[t] > 0$  for  $k \in \mathcal{S}_c[t]$  and  $P_{k,c}[t] = 0$  for  $k \notin \mathcal{S}_c[t]$ .

The ICI power at user  $(k, c)$  receiver in slot  $t$  is given by

$$\chi_{k,c}[t] = \sum_{c' \neq c} g_{k,c,c'} \mathbf{h}_{k,c,c'}^H[t] \Sigma_{c'}[t] \mathbf{h}_{k,c,c'}[t] \quad (3)$$

with mean given by  $\bar{\chi}_{k,c} = \mathbb{E}[\chi_{k,c}[t]] = \sum_{c' \neq c} g_{k,c,c'} \text{tr}(\Sigma_{c'}[t])$ . The SINR at user  $(k, c)$  is given by

$$\text{sinr}_{k,c}[t] = \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^H[t] \mathbf{v}_{k,c}[t]|^2 P_{k,c}[t]}{1 + \chi_{k,c}[t]} \quad (4)$$

We let  $R_{k,c}[t]$  denote the *instantaneous service rate* of user  $(k, c)$  on slot  $t$ , measured in bits/channel use. This is in general a function of  $\text{sinr}_{k,c}[t]$ , and therefore of  $\mathbf{H}_c[t]$ ,  $\Sigma_c[t]$ ,  $\chi_{k,c}[t]$ , and of the allocated rate  $r_{k,c}[t]$ . We define the  $k$ -th user service rate function  $R_k(g, \mathbf{H}, \chi, \Sigma, \mathbf{r})$ , such that  $R_{k,c}[t] = R_k(g_{k,c,c}, \mathbf{H}_c[t], \chi_{k,c}[t], \Sigma_c[t], \mathbf{r}_c[t])$ . Let  $\Gamma$  denote the set of all feasible scheduling policies and let  $\gamma_{-c} = \{\gamma_{c'} : c' \neq c\}$  denote the set of scheduling policies at all cells  $c' \neq c$ . For fixed  $\gamma_{-c} \in \Gamma^{C-1}$ , the throughput of user  $(k, c)$  under the scheduling policy  $\gamma_c$  is given by

$$\begin{aligned} \bar{R}_{k,c}(\gamma_c, \gamma_{-c}) &= \liminf_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t R_k(g_{k,c,c}, \mathbf{H}_c[\tau], \chi_{k,c}[\tau], \gamma_c(\mathbf{H}_c[\tau])) \\ &= \mathbb{E}[R_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \gamma_c(\mathbf{H}_c))] \end{aligned} \quad (5)$$

where the effect of the policies at the interfering BSs is captured by the statistics of the ICI power process  $\chi_{k,c}[t]$ , the limit holds almost surely because of stationarity and ergodicity, and expectation is with respect to the joint distribution of the triple  $(\mathbf{H}_c[t], \chi_{k,c}[t], \gamma_c)$ .<sup>3</sup> The region

<sup>3</sup>With some abuse of notation, we denote by  $\mathbf{H}_c$  and  $\{\chi_{k,c} : k = 1, \dots, K\}$  random variables whose joint distribution coincides with the first-order joint distribution of the processes  $\mathbf{H}_c[t]$  and  $\{\chi_{k,c}[t] : k = 1, \dots, K\}$ , which is time-invariant by stationarity.

of achievable throughputs for cell  $c$  is given by

$$\mathcal{R}_c(\gamma_{-c}) = \text{coh} \bigcup_{\gamma_c \in \Gamma} \left\{ \bar{\mathbf{R}} \in \mathbb{R}_+^K : \bar{R}_k \leq \mathbb{E} [R_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \gamma_c(\mathbf{H}_c))] , \forall k \right\} \quad (6)$$

“coh” denotes “closure of the convex hull”. Notice that  $\mathcal{R}_c(\gamma_{-c})$  depends on the other cells’ scheduling policies  $\gamma_{-c}$  through the joint probability distribution of the ICI powers  $\{\chi_{k,c} : k = 1, \dots, K\}$ .

Under our assumptions, the BSs operate in a decentralized way and influence each other only in terms of the generated ICI statistics (i.e., the joint cdfs  $\{\chi_{k,c} : k = 1, \dots, K\}$ ). Each BS wishes to maximize its own network utility function. This multi-objective optimization problem is formulated as a non-cooperative game [33], [34] that we nickname the *multi-cell decentralized scheduling game*, where each player (i.e., BS)  $c$  seeks to achieve

$$\begin{aligned} & \text{maximize} && U_c(\bar{\mathbf{R}}) \\ & \text{subject to} && \bar{\mathbf{R}} \in \mathcal{R}_c(\gamma_{-c}) \end{aligned} \quad (7)$$

where we assume that  $U_c(\cdot)$  is a continuous, strictly concave and component-wise increasing utility function, reflecting some suitable fairness criterion [6].

By definition, for any given joint statistics of  $\mathbf{H}_c$  and of  $\{\chi_{k,c} : k = 1, \dots, K\}$ , the maximum in (7) is achieved by some scheduling policy  $\gamma_c^*$ . A Nash equilibrium of the *decentralized scheduling game* is a set of scheduling policies (also denoted, with some abuse of notation, by  $\{\gamma_c^* : c = 1, \dots, C\}$ ) such that  $\gamma_c^*$  is the solution to (7) when  $\gamma_{-c} = \gamma_{-c}^*$ , for all  $c = 1, \dots, C$ . We have:

*Theorem 1:* The decentralized scheduling game defined above is a concave game and therefore has a Nash equilibrium.

*Proof:* All players have the same strategy set  $\Gamma$ . This is a compact convex set due to the covariance trace constraint and to the fact that we can assume that the rate allocation vector is bounded in  $\mathbf{r}_c \in [0, r_{\max}]^K$  for some constant  $r_{\max}$ .<sup>4</sup> Also, each  $c$ -th utility is a concave function of  $\gamma_c$  for fixed  $\gamma_{-c}$ . In order to see this, let  $\bar{\mathbf{R}}(\gamma_c, \gamma_{-c})$  denote the throughput point of  $\mathcal{R}_c(\gamma_{-c})$  achieved by policy  $\gamma_c$  for fixed  $\gamma_{-c}$ , consider any two policies  $\gamma'_c, \gamma''_c \in \Gamma$  and define  $\gamma_c^{(\lambda)}$  as the policy that applies  $\gamma'_c$  with probability  $\lambda \in [0, 1]$  and  $\gamma''_c$  with probability  $\bar{\lambda} = 1 - \lambda$ . Then, from

<sup>4</sup>This limitation does not involve any significant loss of generality if  $r_{\max}$  is sufficiently large, and always holds in practice since practical variable-rate coding has a finite maximum rate.

the convexity of  $\mathcal{R}_c(\gamma_{-c})$  and the concavity of  $U_c(\cdot)$  we have that

$$\lambda U_c(\bar{\mathbf{R}}(\gamma'_c, \gamma_{-c})) + \bar{\lambda} U_c(\bar{\mathbf{R}}(\gamma''_c, \gamma_{-c})) \leq U_c(\lambda \bar{\mathbf{R}}(\gamma'_c, \gamma_{-c}) + \bar{\lambda} \bar{\mathbf{R}}(\gamma''_c, \gamma_{-c})) = U_c(\bar{\mathbf{R}}(\gamma_c^{(\lambda)}, \gamma_{-c}))$$

Now, let  $\boldsymbol{\gamma} = \{\gamma_c : c = 1, \dots, C\}$  and  $\boldsymbol{\gamma}' = \{\gamma'_c : c = 1, \dots, C\}$  denote two vectors of scheduling policies and define the sum-utility function  $\rho(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \sum_{c=1}^C U_c(\bar{\mathbf{R}}(\gamma_c, \gamma'_{-c}))$ . Since the functions  $U_c(\cdot)$  are continuous (by assumption) and the throughput vectors are continuous functions of the scheduling policies, it follows that  $\rho(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$  is a continuous function of  $(\boldsymbol{\gamma}, \boldsymbol{\gamma}') \in \Gamma^C \times \Gamma^C$  and, for what said before, it is concave in  $\boldsymbol{\gamma}$  for any fixed  $\boldsymbol{\gamma}'$ . These properties match exactly the assumption of Rosen Theorem [35]. Therefore, as a direct consequence of [35], the existence of a Nash equilibrium is proved. ■

Since  $U_c(\cdot)$  is component-wise increasing, it follows that the maximum of (7) is obtained for some  $\gamma_c^*$  such that  $\bar{\mathbf{R}}(\gamma_c^*, \gamma_{-c})$  is on the Pareto boundary of  $\mathcal{R}_c(\gamma_{-c})$ . If the service rate function  $R_k(g, \mathbf{H}, \chi, \Sigma, \mathbf{r})$  is strictly increasing in the power allocated to user  $k$ , then the Pareto boundary of  $\mathcal{R}_c(\gamma_{-c})$  is achieved by policies that satisfy  $\text{tr}(\Sigma_c[t]) = \sum_{k \in \mathcal{S}_c[t]} P_{k,c}[t] = 1$  with probability 1. In this case, any Nash equilibrium  $\{\gamma_c^* : c = 1, \dots, C\}$  must correspond to scheduling policies that achieve the power constraint with equality for all BSs.

In Sections III and IV we will focus on reference cell  $c$ , assuming that all other interfering cells apply a fixed arbitrary policy  $\gamma_{-c}$  (i.e., for fixed and known joint statistics of the ICI powers at all users of cell  $c$ ). We shall apply the theory developed in [18], [19] and provide a stochastic optimization algorithm that solves (7) to any desired level of approximation, for any given ICI powers statistics.

### III. SCHEDULING WITH ADAPTIVE VARIABLE-RATE CODING AND ARQ-LLC

From now on we shall assume Gaussian random coding and consider specific cases of service rate functions. In this case, we define the  $k$ -th user mutual information function as

$$I_k(g, \mathbf{H}, \chi, \Sigma) = \log \left( 1 + \frac{g |\mathbf{h}_k^\mathsf{H} \mathbf{v}_k|^2 P_k}{1 + \chi} \right) \quad (8)$$

The mutual information at user  $(k, c)$  receiver on slot  $t$  is given by  $I_{k,c}[t] \triangleq I_k(g_{k,c,c}, \mathbf{H}_c[t], \chi_{k,c}[t], \Sigma_c[t])$ . We approximate the decoding error probability by the corresponding *information outage probability* (see [28], [36] for the information-theoretic motivations underlying this very common and very useful approximation). Namely, if the mutual information  $I_{k,c}[t]$  is less than the scheduled

coding rate  $r_{k,c}[t]$ , the decoder makes a decoding error with probability close to 1, while if  $I_{k,c}[t] > r_{k,c}[t]$  the random coding average error probability is very close to 0. Therefore, for slot length  $T$  large enough, there exist “good” codes drawn from a Gaussian ensemble such that their block error probability is close to the information outage probability  $\mathbb{P}(r_{k,c}[t] > I_{k,c}[t])$ . In this case, the user  $k$  service rate function is given by “outage rate” function [18]

$$R_k(g, \mathbf{H}, \chi, \Sigma, \mathbf{r}) = r_k \times \mathbb{1}\{r \leq I_k(g, \mathbf{H}, \chi, \Sigma)\} \quad (9)$$

In order to obtain the desired near-optimal scheduling policy, we apply the framework of [18]. We define the *virtual queues*<sup>5</sup> with buffer state  $\mathbf{Q}_c[t] = (Q_{1,c}[t], \dots, Q_{K,c}[t])$  and *virtual arrival processes*  $\mathbf{A}_c[t] = (A_{1,c}[t], \dots, A_{K,c}[t])$ . The virtual queues evolve according to the stochastic difference equations

$$Q_{k,c}[t+1] = \max\{0, Q_{k,c}[t] - R_{k,c}[t]\} + A_{k,c}[t], \quad k = 1, \dots, K \quad (10)$$

Then, we consider the *adaptive* policy defined by:

- 1) For any given  $t$ , let the transmit covariance matrix  $\Sigma_c[t]$  and the rate allocation vector  $\mathbf{r}_c[t]$  be the solution of

$$\begin{aligned} & \text{maximize} \quad \sum_{k=1}^K Q_{k,c}[t] \mathbb{E}[r_{k,c}[t] \times \mathbb{1}\{r_{k,c}[t] \leq I_k(g_{k,c,c}, \mathbf{H}_c[t], \chi_{k,c}[t], \Sigma_{k,c}[t])\} | \mathbf{H}_c[t]] \\ & \text{subject to} \quad \text{tr}(\Sigma_c[t]) \leq 1, \quad r_{k,c}[t] \geq 0 \quad \forall k \end{aligned} \quad (11)$$

- 2) For suitable constants  $V, A_{\max} > 0$ , let the virtual arrival processes at time  $t$  be given by the solution of

$$\max_{0 \leq A_{k,c}[t] \leq A_{\max}, \forall k} VU_c(\mathbf{A}_c[t]) - \sum_{k=1}^K A_{k,c}[t] Q_{k,c}[t] \quad (12)$$

- 3) Update the virtual queues according to (10), with arrivals  $\mathbf{A}_c[t]$  given by (12) and service rates  $R_{k,c}[t]$  given by (9) calculated for  $\Sigma_c[t]$  and  $\mathbf{r}_c[t]$  solutions of (11).

As stated in Theorem 2 below, the policy defined above achieves the optimal point  $\bar{\mathbf{R}}_c^*$  solution of (7) within any desired accuracy, depending on the constants  $V$  and  $A_{\max}$ . Neglecting the (small) degradation due to stochastic adaptation and quantified by Theorem 2, we shall refer to this policy as  $\gamma^*$ .

<sup>5</sup>It is important to keep in mind that the virtual queues have nothing to do with the ARQ transmission buffers: they are used here as a tool to recursively update the weights of the the scheduling policy.

*Theorem 2:* Assume i.i.d. channels and fixed joint statistics of the ICI powers  $\{\chi_{k,c} : k = 1, \dots, K\}$ . Assume that  $U_c(\cdot)$  is concave and entry-wise non-decreasing, and that there exists at least one point  $\mathbf{r} \in \mathcal{R}_c(\gamma_{-c})$  with strictly positive entries such that  $U_c(\mathbf{r}/2) > -\infty$ . Then, the scheduling policy  $\gamma_c^*$  defined above, for given constants  $V > 0$  and  $A_{\max} > 0$ , has the following properties:

- (a) The utility achieved by  $\gamma^*$  satisfies:

$$\liminf_{t \rightarrow \infty} U_c \left( \frac{1}{t} \sum_{\tau=1}^t \mathbf{R}_c[\tau] \right) \geq U_c(\bar{\mathbf{R}}_c^*(A_{\max})) - \kappa/V \quad (13)$$

where

$$\kappa \triangleq \frac{1}{2} \left( K A_{\max}^2 + \sum_{k=1}^K \mathbb{E} \left[ \log^2 \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}|^2}{1 + \chi_{k,c}} \right) \right] \right) \quad (14)$$

and where  $\bar{\mathbf{R}}_c^*(A_{\max})$  denotes the solution of the problem (7) with the additional constraint  $0 \leq \bar{R}_{k,c} \leq A_{\max}$  for all  $k = 1, \dots, K$ .

- (b) For any point  $\bar{\mathbf{R}}_c \in \mathcal{R}_c(\gamma_{c'} : c' \neq c)$  such that  $0 \leq \bar{R}_{k,c} \leq A_{\max}$  for all  $k$ , and for any value  $\beta \in [0, 1]$  we have:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t \sum_{k=1}^K \bar{R}_{k,c} \mathbb{E}[Q_{k,c}[\tau]] \leq \frac{\kappa + V[U_c(\bar{\mathbf{R}}_c^*(A_{\max})) - U_c(\beta \bar{\mathbf{R}}_c)]}{1 - \beta} \quad (15)$$

Thus, all virtual queues  $Q_{k,c}[t]$  are strongly stable.<sup>6</sup>

*Proof:* The proof follows verbatim from the results in [18] and it is not repeated here for brevity. ■

As a corollary of Theorem 2, if  $A_{\max}$  is sufficiently large such that  $A_{\max} \geq \bar{R}_{k,c}^*$  for all  $k$ , then  $\gamma_c^*$  satisfies

$$\liminf_{t \rightarrow \infty} U_c \left( \frac{1}{t} \sum_{\tau=1}^t \mathbf{R}_c[\tau] \right) \geq U_c(\bar{\mathbf{R}}_c^*) - \kappa/V. \quad (16)$$

Hence, the control parameter  $V$  can be chosen sufficiently large in order to make the achieved utility as close as desired to the optimal value  $U_c(\bar{\mathbf{R}}_c^*)$  of problem (7). This comes with a tradeoff in the virtual queue average sizes that, as seen from (15), grow linearly with  $V$ . The virtual queue sizes represent the difference between the virtual bits admitted into the queues and the actual bits transmitted, and thus affect the time-scales required for the time averages to become close to their limiting values.

<sup>6</sup> A discrete-time queue  $Q_k[t]$  is *strongly stable* if  $\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t \mathbb{E}[Q_k[\tau]] < \infty$ . The system is strongly stable if all queues  $k = 1, \dots, K$  are strongly stable.

### A. Implementation

The policy  $\gamma^*$  found before computes recursively the “weights”  $\mathbf{Q}_c[t]$  via (12) and (10) and, for each  $t$ , solves the weighted conditional average rate sum maximization (11). Problem (12) is a standard convex optimization problem the solution of which does not present any major conceptual difficulty and is found in closed form for the important cases of proportional fairness and max-min fairness (see [18]), corresponding to the choices  $U_c(\overline{\mathbf{R}}) = \sum_{k=1}^K \log \overline{R}_k$  and  $U_c(\overline{\mathbf{R}}) = \min_k \overline{R}_k$ , respectively. In contrast, solving (11) presents some difficulties. Letting  $F_{k,c}(\cdot)$  denote the marginal cdf of  $\chi_{k,c}[t]$  and using (8), the objective function in (11) can be rewritten as

$$\sum_{k \in \mathcal{S}_c[t]} Q_{k,c}[t] r_{k,c}[t] F_{k,c} \left( \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\mathsf{H}[\mathbf{v}_{k,c}[t]]|^2 P_{k,c}[t]}{2^{r_{k,c}[t]} - 1} - 1 \right) \quad (17)$$

The optimization in (11) is generally a non-convex problem that involves a combinatorial search over all subsets  $\mathcal{S}_c[t] \subseteq \{1, \dots, K\}$  of cardinality  $\leq \min\{K, M\}$  and, for each candidate subset, the maximization of (17) with respect to  $\mathbf{r}_c[t]$  and the power allocation  $\{P_{k,c}[t] : k \in \mathcal{S}_c[t]\}$ . Since this optimization may be difficult to compute, we propose the following suboptimal low-complexity two-step approach:

Step 1) the active user subset and the corresponding power allocation are selected by assuming deterministic ICI powers, equal to their mean value  $\overline{\chi}_{k,c}$ . Under this assumption, the problem is reduced to the well-known user selection with ZFBF, that can be solved using standard techniques based on quasi-orthogonal user selection and waterfilling (e.g., [37]–[39]).

Step 2) for the transmit covariance  $\Sigma_c[t]$  obtained in step 1, (17) is optimized with respect to the rate allocation. This reduces to optimizing the outage rate separately for each  $k \in \mathcal{S}_{k,c}[t]$  by letting

$$r_{k,c}[t] = \arg \max_{r \geq 0} \left\{ r F_{k,c} \left( \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\mathsf{H}[\mathbf{v}_{k,c}[t]]|^2 P_{k,c}[t]}{2^r - 1} - 1 \right) \right\} \quad (18)$$

where  $g_{k,c,c} |\mathbf{h}_{k,c,c}^\mathsf{H}[\mathbf{v}_{k,c}[t]]|^2 P_{k,c}[t]$  is fixed by Step 1.

Notice that, both in the original problem and in the proposed low-complexity two-step approximated solution, only the *marginal* statistics of the ICI powers  $\{\chi_{k,c}[t] : k = 1, \dots, K\}$  are relevant. These marginal statistics can be measured by each user terminal individually and fed back to the BS scheduler by some very low-rate feedback scheme.

#### IV. SCHEDULING WITH INCREMENTAL REDUNDANCY HARQ

If a genie provides the BS scheduler with the values of the the mutual information  $\{I_{k,c}[t] : k = 1, \dots, K\}$  in a non-causal fashion, just before the beginning of slot  $t$ , then the optimal rate allocation would be, trivially,  $r_{k,c}[t] = I_{k,c}[t]$  for all  $k \in \mathcal{S}_c[t]$ , yielding zero outage probability. This “genie-aided” case was considered in [18] and referred to as “optimistic rate” allocation, although no actual algorithm to approach the optimistic throughput was given. Since for any non-negative random variable  $I$  and  $r > 0$  we have  $\mathbb{E}[r \mathbb{1}\{r > I\}] \leq \mathbb{E}[I]$ , then the optimistic service rates provide an upper bound to the throughput of any system with the same signaling scheme (ZFBF and Gaussian codes) and given rate allocation.

In this section we show how to achieve the “optimistic” throughput without the aid of any genie. As a preliminary step, let’s consider the following *incremental redundancy* HARQ scheme. The BS scheduler maintains a buffer of information packets for each user in the cell. The size of user  $(k, c)$  packets is equal to  $b_{k,c}$  bits per packet. Each packet is encoded into an infinite-length sequence of complex symbols.<sup>7</sup> The encoded sequence is partitioned into blocks of length  $T$  symbols. At each slot  $t$ , the scheduling policy computes  $\Sigma_c[t]$  according to some rule to be found later. For all active users  $k \in \mathcal{S}_c[t]$ , if the most recent HARQ feedback message from user  $k$  is “NACK” (negative acknowledgement), then the first not-yet transmitted coded block of the current packet is transmitted. Otherwise, if the most recent received HARQ feedback message is “ACK” (positive acknowledgement), then the current packet is removed from the transmission buffer of user  $k$  and the first coded block of next packet in the buffer is transmitted. The  $(k, c)$ -th receiver stores in memory all the received slots for times  $\{t : k \in \mathcal{S}_c[t]\}$  and attempts to decode the current packet at every newly received slot, by using all the available received slots. If decoding fails, NACK is sent back, otherwise ACK is sent back and the decoder memory is reset. Notice that the scheme does not require any genie-aided “look-ahead” of the instantaneous ICI power  $\chi_{k,c}[t]$ , and makes use of time-invariant packet sizes  $b_{k,c}$ . These may differ from user to user but are independent of  $t$ . For later use, we define the “first-block coding rate” as the ratio  $r_{k,c} = \frac{b_{k,c}}{T}$  bits/channel use.

Next, we describe a scheduling rule, denoted again by  $\gamma_c^*$ , that operates arbitrarily closely to the genie-aided throughput when combined with the HARQ scheme described above. At the

<sup>7</sup>In practice, this rateless coding can be implemented by using Raptor codes [40].

end of each slot  $t$ , the active users  $k \in \mathcal{S}_c[t]$  feed back both their ACK/NACK message and the mutual information  $I_{k,c}[t]$  “seen” at their receiver. Then,  $\gamma_c^*$  coincides with what given in Section III, after the following two changes. 1) The virtual queues evolution equation (10) is replaced by

$$Q_{k,c}[t+1] = \max \{0, Q_{k,c}[t] - I_{k,c}[t]\} + A_{k,c}[t], \quad \forall k \quad (19)$$

2) The transmitter optimization (11) is replaced by

$$\begin{aligned} & \text{maximize} \quad \sum_{k=1}^K Q_{k,c}[t] \mathbb{E} [I_k(g_{k,c,c}, \mathbf{H}_c[t], \chi_{k,c}[t], \Sigma_c[t]) | \mathbf{H}_c[t]] \\ & \text{subject to} \quad \text{tr}(\Sigma_c[t]) \leq 1 \end{aligned} \quad (20)$$

In brief, the scheduler updates recursively its weights  $\mathbf{Q}_c[t]$  and computes the transmitted signal covariance  $\Sigma_c[t]$  according to (20), as if it was operating on a virtual “genie-aided” system with instantaneous service rates  $I_{k,c}[t]$ . The throughput region of the virtual genie-aided system, denoted by  $\mathcal{R}_c^{\text{genie}}(\gamma_{-c})$ , is given by (6), after replacing the general rate function  $R_k(\dots)$  with the mutual information function  $I_k(\dots)$  defined in (8). The performance of  $\gamma_c^*$  for the genie-aided system is again given by Theorem 2, where  $\mathbf{R}_c[\tau]$  in (13) is replaced by the vector of mutual informations  $\mathbf{I}_c[\tau] = (I_{1,c}[\tau], \dots, I_{K,c}[\tau])$  and where  $\overline{\mathbf{R}}^*(A_{\max})$  denotes the solution of (7) when  $\mathcal{R}_c(\gamma_{-c})$  is replaced by  $\mathcal{R}_c^{\text{genie}}(\gamma_{-c})$ , with the additional constraint  $0 \leq \overline{R}_{k,c} \leq A_{\max}$  for all  $k = 1, \dots, K$ .

For sufficiently large  $A_{\max}$ ,  $\gamma_c^*$  yields:

$$\liminf_{t \rightarrow \infty} U_c \left( \frac{1}{t} \sum_{\tau=1}^t \mathbf{I}_c[\tau] \right) \geq U_c(\overline{\mathbf{R}}_c^{\text{genie},*}) - \kappa/V, \quad (21)$$

where  $\overline{\mathbf{R}}_c^{\text{genie},*}$  is the utility-maximizing throughput point in the region  $\mathcal{R}_c^{\text{genie}}(\gamma_{-c})$ . At this point, it remains to be shown that the combination of the policy  $\gamma_c^*$  with the incremental redundancy HARQ scheme yields a network utility as close as desired to the limit in (21). This is shown by the following:

*Theorem 3:* Let  $\overline{\mathbf{R}}_c^{\text{harq},*} = (\overline{R}_{1,c}^{\text{harq},*}, \dots, \overline{R}_{K,c}^{\text{harq},*})$  denote the throughput achievable by the incremental redundancy HARQ protocol under scheduling policy  $\gamma_c^*$  defined above. For each user  $(k, c)$  and  $\epsilon_{k,c} > 0$  there exists a sufficiently large first-block rate  $r_{k,c}$  such that  $\overline{R}_{k,c}^{\text{harq},*} \geq (1 - \epsilon_{k,c}) \overline{R}_{k,c}^{\text{genie},*}$ .

*Proof:* Consider user  $(k, c)$ . Following the argument in [27], we can model the event of successful decoding as a “mutual information level-crossing event”. Suppose that the transmission of the current packet for user  $(k, c)$  starts at slot  $t_{\text{start}}$  (i.e., an ACK was fed back at slot time  $t_{\text{start}} - 1$ ). Then, the current packet can be successfully decoded at slot  $t \geq t_{\text{start}}$  if  $\sum_{\tau=t_{\text{start}}}^t I_{k,c}[\tau] \geq r_{k,c}$ . Otherwise, a decoding error occurs with very high probability. As shown in [27], [41], the probability of undetected decoding error vanishes exponentially with  $T$ . Therefore, in the regime of large  $T$ , if  $\sum_{\tau=t_{\text{start}}}^t I_{k,c}[\tau] < r_{k,c}$  the decoding error is detected with arbitrarily high probability and a NACK is sent back. Fig. 1 shows, qualitatively, the mutual information level-crossing and the corresponding successful decoding events of the  $(k, c)$  decoder. Notice that the mutual information increment is non-negative, and it is exactly zero for all  $t$  such that  $k \notin \mathcal{S}_c[t]$ , i.e., when user  $(k, c)$  is not scheduled.

Consider the transmission of a long sequence of packets. Without loss of generality, assume that the system starts at time  $t_{\text{start}} = 1$ , denote by  $N_{k,c}[t]$  the number of successful decoding events of decoder  $(k, c)$  up to time  $t$  and let  $W_{k,c}(1), W_{k,c}(2), \dots, W_{k,c}(N_{k,c}[t])$  denote the corresponding “inter-ACK” times (see Fig. 1). Since at each successful decoding a “reward” of  $r_{k,c}$  bit per channel use is delivered to the destination, the throughput of the HARQ protocols is given by

$$\overline{R}_{k,c}^{\text{harq},*} = \lim_{t \rightarrow \infty} \frac{r_{k,c} N_{k,c}[t]}{\sum_{n=1}^{N_{k,c}[t]} W_{k,c}(n) + \Delta_{k,c}[t]} \quad (22)$$

where  $\Delta_{k,c}[t] = t - \sum_{n=1}^{N_{k,c}[t]} W_{k,c}(n)$  denotes the difference between the current time  $t$  and the time at which the  $N_{k,c}[t]$ -th successful decoding occurred. Under the assumptions of this paper, the system with HARQ protocol and scheduling policy  $\gamma_c^*$  evolves according to a discrete-time, continuous-valued vector Markov process with state given by  $\mathbf{Q}_c[t]$  and by the vector of accumulated mutual informations at each receiver. Since the virtual queues are strongly stable (see Theorem 2) and the accumulated mutual informations are bounded in  $[0, r_{k,c}]$ , the process is stationary and ergodic. Therefore, the limit in (22) holds almost surely, and can be explicitly

computed as follows:

$$\begin{aligned}
\bar{R}_{k,c}^{\text{harq},\star} &= \lim_{t \rightarrow \infty} \frac{r_{k,c}}{\frac{1}{N_{k,c}[t]} \sum_{n=1}^{N_{k,c}[t]} W_{k,c}(n) + \frac{\Delta_{k,c}[t]}{N_{k,c}[t]}} \\
&= \frac{r_{k,c}}{\lim_{t \rightarrow \infty} \frac{1}{N_{k,c}[t]} \sum_{n=1}^{N_{k,c}[t]} W_{k,c}(n) + \lim_{t \rightarrow \infty} \frac{\Delta_{k,c}[t]}{N_{k,c}[t]}} \\
&= \frac{r_{k,c}}{\mathbb{E}[W_{k,c}]}
\end{aligned} \tag{23}$$

where  $W_{k,c}$  is an integer-valued random variable with the same marginal distribution of the inter-ACK times.

In order to determine  $\mathbb{E}[W_{k,c}]$ , consider the case  $t_{\text{start}} = 1$  and define the event

$$\mathcal{A}_{k,c}[t] = \left\{ \sum_{\tau=1}^t I_k[\tau] \leq r_{k,c} \right\} \tag{24}$$

Since the accumulated mutual information between two ACKs is non-decreasing, the following nesting condition holds:

$$\mathcal{A}_{k,c}[t] \subseteq \mathcal{A}_{k,c}[t-1], \quad \forall t$$

where  $\mathcal{A}_{k,c}[0] = \{0 \leq r_{k,c}\}$  has probability 1. It follows that

$$\mathbb{P}(W_{k,c} = t) = \mathbb{P}(\mathcal{A}_{k,c}[t-1], \overline{\mathcal{A}_{k,c}[t]}) = \mathbb{P}(\mathcal{A}_{k,c}[t-1]) - \mathbb{P}(\mathcal{A}_{k,c}[t]),$$

yielding the average inter-ACK time in the form

$$\begin{aligned}
\mathbb{E}[W_{k,c}] &= \sum_{t=1}^{\infty} t \mathbb{P}(W_{k,c} = t) \\
&= 1 + \sum_{t=1}^{\infty} \mathbb{P}(\mathcal{A}_{k,c}[t])
\end{aligned} \tag{25}$$

Owing to the complete formal analogy of results (23) and (25) with the throughput of HARQ considered in [27]), we can directly apply the limit proved in [27]:<sup>8</sup>

$$\lim_{r_{k,c} \rightarrow \infty} \frac{r_{k,c}}{\mathbb{E}[W_{k,c}]} = \mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \Sigma_c)] \tag{26}$$

<sup>8</sup>This result is indeed quite intuitive: when  $r_{k,c}$  becomes large, then  $\mathbb{E}[W_{k,c}]$  increases. Therefore, the accumulated mutual information divided by the number of slots  $\frac{1}{W_{k,c}} \sum_{\tau=1}^{W_{k,c}} I_k(g_{k,c,c}, \mathbf{H}_c[\tau], \chi_{k,c}[\tau], \Sigma_c[\tau])$  converges to an ensemble average. It follows that in this limit the level crossing condition tends to become deterministic, and satisfies (approximately)

$$\sum_{\tau=1}^{W_{k,c}} I_k(g_{k,c,c}, \mathbf{H}_c[\tau], \chi_{k,c}[\tau], \Sigma_c[\tau]) = W_{k,c} r_{k,c}$$

Of course, this argument can be made rigorous by following in the footsteps of [27].

In particular, as  $r_{k,c} \rightarrow \infty$  the average inter-ACK time  $\mathbb{E}[W_{k,c}]$  diverges to infinity linearly with  $r_{k,c}$ . The analysis in [27] shows that, for any  $\eta_{k,c} > 0$ ,

$$\overline{R}_{k,c}^{\text{harq},\star} \geq (1 - \eta_{k,c}) \mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \Sigma_c)] \quad (27)$$

for all sufficiently large  $r_{k,c}$ .

The proof of Lemma 3 is finally concluded by combining the result (26) with (21). By stationarity and ergodicity, under  $\gamma_c^*$  we have that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t I_{k,c}[\tau] = \mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \Sigma_c)]$$

holds almost surely. Since  $U_c(\cdot)$  is component-wise increasing, (21) implies that for any  $\delta_{k,c} > 0$  there exist sufficiently large  $A_{\max}$  and  $V$  for which

$$\mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \Sigma_c)] \geq (1 - \delta_{k,c}) \overline{R}_{k,c}^{\text{genie},\star} \quad (28)$$

By letting  $(1 - \epsilon_{k,c}) = (1 - \eta_{k,c})(1 - \delta_{k,c})$  and using (27) and (28) Theorem 3 is proved. ■

From the above proof it follows that the delay-throughput operating point of the incremental redundancy HARQ protocol can be chosen individually for each user by setting the threshold value  $r_{k,c}$  (or, equivalently, the size  $b_{k,c}$  of the information packets). By making  $r_{k,c}$  large, the average decoding delay  $D_{k,c} = \mathbb{E}[W_{k,c}]$  becomes large and the throughput approaches  $\overline{R}_{k,c}^{\text{genie},\star}$ .

Also, we wish to stress the difference between the ARQ-LLC scheme described in Section II and the incremental-redundancy HARQ protocol illustrated in this section. The ARQ-LLC protocol makes use of adaptive variable-rate coding at the physical layer, and removes or keeps in the transmission buffer packets of information bits of variable size  $b_{k,c}[t] = Tr_{k,c}[t]$ . In contrast, the HARQ protocol make use of a fixed packet size  $b_{k,c}$  (equivalent to fixed first-block rate  $r_{k,c}$ ), but the effective service rate is adaptive by varying the decoding delay through the ACK/NACK mechanism.

### A. Implementation

The scheme previously proposed requires that each active user, at the end of each slot  $t$ , feeds back a message formed by one bit for ACK/NACK and by the value of  $I_{k,c}[t]$  or, equivalently, the value of  $\text{sinr}_{k,c}[t]$ . We notice that feeding back the instantaneous SINR is widely proposed in the literature on opportunistic downlink scheduling [42], [43] and it is referred to as *Channel*

*Quality Indicator* (CQI). However, in the current literature the CQI is relative to the *current* slot, and it is used to select users and allocate the rate of a variable-rate coding scheme. In contrast, here the CQI refers to the *past* slot, and it is used to update the scheduler weights according to (19).

Denoting again by  $F_{k,c}(\cdot)$  the marginal cdf of  $\chi_{k,c}[t]$ , the objective function in (20) can be rewritten as

$$\sum_{k \in \mathcal{S}_c[t]} Q_{k,c}[t] \int_0^\infty \log \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\text{H}[\mathbf{v}_{k,c}[t]]|^2 P_{k,c}[t]}{1+z} \right) dF_{k,c}(z) \quad (29)$$

While for any fixed user subset  $\mathcal{S}_c[t]$  the maximization of (29) with respect to the powers  $\{P_{k,c}[t] : k \in \mathcal{S}_c[t]\}$  is a convex problem, the solution is not generally given by the simple waterfilling formula and it may be difficult to compute since the cdfs  $F_{k,c}(\cdot)$  are typically not known in closed form. A near-optimum low-complexity approximation consists of choosing  $\Sigma_c[t]$  that maximizes the objective function lower bound

$$\sum_{k \in \mathcal{S}_c[t]} Q_{k,c}[t] \log \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\text{H}[\mathbf{v}_{k,c}[t]]|^2 P_{k,c}[t]}{1+\bar{\chi}_{k,c}} \right) \quad (30)$$

obtained by applying Jensen's inequality to (29). Notice that the maximization of (30) with respect to the transmit covariance matrix coincides with step 1 in the low-complexity approximation of the variable-rate coding/ARQ-LLC case of Section III-A, and can be solved efficiently using the methods in [37]–[39].

### B. Extremal ICI distributions

The throughput performance of the HARQ scheme depends on the statistics of the ICI powers, which in turns depend on the scheduling policies  $\gamma_{-c}$  at the interfering BSs. In this section we find *extremal* marginal statistics for the ICI powers that provide non-trivial inner and outer bounds to  $\mathcal{R}_c^{\text{genie}}(\gamma_{-c})$  that are independent of  $\gamma_{-c}$ . Here we drop the slot index  $t$  since all processes are stationary. We start with the following:

*Lemma 1:* For all feasible policies  $\gamma_{c'} : c' \neq c$  that satisfy the input power constraint with equality and for all users  $k = 1, \dots, K$ , we have

$$\mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \bar{\chi}_{k,c}, \Sigma_c)] \leq \mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \chi_{k,c}, \Sigma_c)] \leq \mathbb{E}[I_k(g_{k,c,c}, \mathbf{H}_c, \tilde{\chi}_{k,c}, \Sigma_c)] \quad (31)$$

where  $\bar{\chi}_{k,c} = \mathbb{E}[\chi_{k,c}] = \sum_{c' \neq c} g_{k,c,c'}$  and where  $\tilde{\chi}_{k,c} = \sum_{c' \neq c} g_{k,c,c'} |\mathbf{h}_{k,c,c'}^\text{H} \mathbf{v}_{1,c'}|^2$  is the ICI power at the  $(k, c)$  receiver when all interfering BSs  $c' \neq c'$  schedule a single user in their own cell.

*Proof:* The first inequality (lower bound) follows immediately from Jensen's inequality applied to the convex function  $f(x) = \log(1 + \frac{a}{b+x})$  with  $a, b > 0$ , and by the fact that, by assumption, the interfering BSs use all their available power. In order to show the second inequality (upper bound), we use (2) in (3) and write  $\chi_{k,c} = \sum_{c' \neq c} g_{k,c,c'} \sum_{j \in \mathcal{S}_{c'}} \alpha_{k,c,c',j} P_{j,c'}$ , where  $\alpha_{k,c,c',j} \triangleq |\mathbf{h}_{k,c,c'}^\text{H} \mathbf{v}_{j,c'}|^2$  are random variables independent of the SINR numerator  $|\mathbf{h}_{k,c,c'}^\text{H} \mathbf{v}_{k,c}|^2 P_{k,c}$ . Since the ZFBF steering vectors  $\mathbf{v}_{j,c'}$  have unit norm and are independent of  $\mathbf{h}_{k,c,c'}$ , the variables  $\alpha_{k,c,c',j}$  are marginally identically distributed as central chi-squared with 2 degrees of freedom [44]. Also, notice that the  $\alpha_{k,c,c',j}$ 's are statistically dependent for the same index  $c'$ , while  $\{\alpha_{k,c,c',j} : j \in \mathcal{S}_{c'}\}$  and  $\{\alpha_{k,c,c'',j} : j \in \mathcal{S}_{c''}\}$  are group-wise mutually independent for  $c' \neq c''$ . By assumption,  $\sum_{j \in \mathcal{S}_{c'}} P_{j,c'} = 1$  for all  $c'$ . Therefore,  $\sum_{j \in \mathcal{S}_{c'}} \alpha_{k,c,c',j} P_{j,c'}$  is a convex combination of identically distributed, possibly dependent, random variables. The second inequality in (31) follows by repeated application of Jensen's inequality. Choose  $c'' \neq c$ . Then, using (8), we have

$$\begin{aligned} & \mathbb{E} \left[ \log \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\text{H} \mathbf{v}_{k,c}|^2 P_{k,c}}{1 + \chi_{k,c}} \right) \right] \leq \\ & \sum_{j \in \mathcal{S}_{c''}} P_{j,c''} \mathbb{E} \left[ \log \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\text{H} \mathbf{v}_{k,c}|^2 P_{k,c}}{1 + g_{k,c,c''} \alpha_{k,c,c'',j} + \sum_{c' \neq c,c''} g_{k,c,c'} \sum_{j \in \mathcal{S}_{c'}} \alpha_{k,c,c',j} P_{j,c'}} \right) \right] = \\ & \mathbb{E} \left[ \log \left( 1 + \frac{g_{k,c,c} |\mathbf{h}_{k,c,c}^\text{H} \mathbf{v}_{k,c}|^2 P_{k,c}}{1 + g_{k,c,c''} \alpha_{k,c,c'',1} + \sum_{c' \neq c,c''} g_{k,c,c'} \sum_{j \in \mathcal{S}_{c'}} \alpha_{k,c,c',j} P_{j,c'}} \right) \right] \end{aligned} \quad (32)$$

where the equality in (32) follows from the fact that the  $\alpha_{k,c,c'',j}$ 's are identically distributed with respect to the index  $j$ . Next, pick  $c''' \neq c, c''$ , and apply the same steps to the last line of (32). After eliminating all convex combinations, the final upper bound coincides with the right most term in (31). ■

As a corollary, we have the following interesting “robustness” result:

*Theorem 4:* For any choice of the scheduling policies  $\gamma_{-c}$  that satisfy the input power constraint with equality, we have

$$\overline{\mathcal{R}}_c \subseteq \mathcal{R}_c^{\text{genie}}(\gamma_{-c}) \subseteq \widetilde{\mathcal{R}}_c^{\text{genie}} \quad (33)$$

where  $\overline{\mathcal{R}}_c$  is the region with deterministic ICI powers  $\{\overline{\chi}_{k,c}\}$ ,<sup>9</sup> and where  $\widetilde{\mathcal{R}}_c^{\text{genie}}$  is the region

<sup>9</sup>Notice that if the ICI powers were deterministic, then no genie or HARQ is needed and the system reduces to a collection of isolated cells, where each cell  $c$  has modified channel path gain coefficients  $\overline{g}_{k,c,c} = \frac{g_{k,c,c}}{1 + \overline{\chi}_{k,c}}$ . In this case, the throughput region  $\overline{\mathcal{R}}_c$  is achieved by the standard scheduling/resource allocation schemes with perfect state information and zero outage probability.

corresponding to random ICI powers  $\{\tilde{\chi}_{k,c}\}$ . Furthermore, the gap between the inner and outer bounds in (33) is bounded by a constant that does not depend on the channel path coefficients.

*Proof:* The proof (33) follows directly as a consequence of Lemma 1. In order to show the bounded gap, we have to find some constant  $\Delta$ , independent of  $\{g_{k,c,c'}\}$ , such that  $\max\{\mathbf{r} - \Delta\mathbf{1}, \mathbf{0}\} \in \overline{\mathcal{R}}_c$  for all points  $\mathbf{r} \in \widetilde{\mathcal{R}}_c^{\text{genie}}$ . To this purpose, pick a point  $\mathbf{r} \in \widetilde{\mathcal{R}}_c^{\text{genie}}$  corresponding to some feasible scheduling policy  $\gamma_c$  for the genie-aided system. Applying the *same* sequence of input covariance matrices as determined by  $\gamma_c$ , to the system with deterministic ICI powers, we certainly find a point  $\overline{\mathbf{R}}_c(\gamma_c) \in \overline{\mathcal{R}}_c$ . Consider the throughput of the  $k$ -th user and let for convenience  $A \triangleq g_{k,c,c} |\mathbf{h}_{k,c,c}^H \mathbf{v}_{k,c}|^2 P_{k,c}$ . Then, by applying Jensen's inequality we have

$$\begin{aligned} & \mathbb{E} \left[ \log \left( 1 + \frac{A}{1 + \sum_{c' \neq c} g_{k,c,c'} \alpha_{k,c,c',1}} \right) \middle| A \right] - \log \left( 1 + \frac{A}{1 + \sum_{c' \neq c} g_{k,c,c'}} \right) \\ & \leq \\ & \log \left( 1 + \sum_{c' \neq c} g_{k,c,c'} \right) - \mathbb{E} \left[ \log \left( 1 + \sum_{c' \neq c} g_{k,c,c'} \alpha_{k,c,c',1} \right) \right] \end{aligned} \quad (34)$$

The RHS in the above inequality is easily seen to be non-negative and component-wise increasing with respect to any coefficient  $g_{k,c,c'}$ . Therefore, its maximum is obtained in the limit for all  $g_{k,c,c'} \rightarrow \infty$  (in passing, we notice that this corresponds to considering the interference-limited regime where  $\text{SNR} \rightarrow \infty$ ). In order to see that this limit is finite, let  $g_{\max} = \max g_{k,c,c'}$ , then we have

$$\begin{aligned} \text{RHS of (34)} & \leq \log(1 + (C-1)g_{\max}) - \mathbb{E} \left[ \log \left( 1 + g_{\max} \sum_{c' \neq c} \alpha_{k,c,c',1} \right) \right] \\ & \leq -\mathbb{E} \left[ \log \left( \frac{1}{C-1} \sum_{c' \neq c} \alpha_{k,c,c',1} \right) \right] \end{aligned} \quad (35)$$

$$\leq -\mathbb{E} [\log(\alpha_{k,c,c',1})] \quad (36)$$

$$\leq \gamma / \ln(2) \quad (37)$$

where (35) follows by letting  $g_{\max} \rightarrow \infty$ , (36) follows by applying Jensen's inequality to the convex function  $-\log x$  and (37) follows by using the fact that  $\alpha_{k,c,c',1}$  is chi-squared with 2 degrees of freedom, and using the limit  $\lim_{\epsilon \downarrow 0} \int_{\epsilon}^{\infty} \ln x e^{-x} dx = -\gamma$ , where  $\gamma$  denotes the Euler-Mascheroni constant [45]. ■

Theorem 4 has the following interesting consequence: consider the multi-cell decentralized scheduling game under the proposed incremental redundancy HARQ scheme, achieving the

genie-aided throughput region in each cell. The performance of any given cell  $c$  (in terms of its network utility value) at any Nash equilibrium  $(\gamma_1^*, \dots, \gamma_C^*)$  is bounded below and above by the solutions of (7) when  $\mathcal{R}_c^{\text{genie}}(\gamma_{-c}^*)$  is replaced by  $\overline{\mathcal{R}}_c$  and  $\widetilde{\mathcal{R}}_c^{\text{genie}}$ , respectively. This follows from the fact that, as argued at the end of Section II, all Nash equilibria must achieve the power constraints with equality at each BS.<sup>10</sup>

## V. NUMERICAL RESULTS

We considered a simple one-dimensional cellular layout with unit width cells arranged on a line. BSs are located at integer positions  $c \in \mathbb{Z}$ . In each cell  $c$ , users are placed on a uniform grid in positions  $u(k, c) = (2k - K - 1)/(2K) + c$ , for  $k = 1, \dots, K$ . The channel path gains are given by  $g_{k,c,c'} = \frac{G_0}{1 + (|u(k,c) - c'|_C/\delta)^\nu}$ , where the modulo- $C$  distance  $|u - c|_C = \min\{|u - c + zC| : z \in \mathbb{Z}\}$  induces a torus topology that eliminates border effects and where  $\nu$  and  $\delta$  are the propagation exponent and the 3dB breakpoint distance, respectively, and  $G_0$  determines the received SNR at the cell edge [29]. We present results for a system with  $C = 18$  cells,  $M = 2$  antennas per BS,  $K = 36$  users per cell and parameters  $G_0 = 60\text{dB}$ ,  $\alpha = 3.0$  and  $\delta = 0.05$ . For the implementation of the policy  $\gamma_c^*$  we chose parameters  $A_{\max} = 50$ ,  $V = 50$  and suboptimal low-complexity approximations as explained in Sections III-A and IV-A, respectively. As for the network utility functions, we considered both proportional fairness and max-min fairness (see Section III-A and [6], [18] and references therein). In order to gather the ICI statistics, we run the same scheduling algorithm in all BSs and measure the empirical cdf of the ICI power at each user location in the reference cell  $c = 0$  (since the system is completely symmetric, all cells see the same ICI statistics).

Figs. 2 and 3 compare user throughputs in cell  $c = 0$  under proportional fairness and max-min fairness, respectively. Thick dashed lines illustrate the throughput upper bounds of Theorem 4. Thin dashed lines correspond to the actual “genie-aided” rates achievable by the proposed HARQ scheme in the limit of infinite decoding delay. Solid lines show the throughput achieved by the HARQ scheme operating at finite average decoding delay for all users, by setting the parameters  $\{r_{k,0}\}$  such that each user achieves 97% of the genie-aided rates (infinite delay). The “triangle” marks indicate the throughput lower bounds of Theorem 4. Finally, the “square” marks indicate

<sup>10</sup>Notice that the mutual information function is strictly increasing with the SINR.

the throughputs achieved by the conventional adaptive variable-rate coding with ARQ-LLC. We observe that under both fairness objective functions, the throughputs achieved by HARQ achieve a gain of more than 100% for the users at the edge of the cell in the proportional fairness case, and a throughput gain of more than 40% for all users in the max-min fairness case, with respect to the ARQ-LLC scheme.

Figs. 4 and 5 illustrate the average throughput as a function of the average decoding delay for the HARQ scheme in the case of two specific users: user  $(1, 0)$  at the left cell edge and  $(18, 0)$  at the cell center, under proportional fairness and max-min fairness, respectively. The thick dashed lines show genie-aided rates. The solid lines are obtained by increasing first-block coding rate parameter  $r_{k,0}$  and computing average decoding delay from (25) with  $\mathbb{P}(\mathcal{A}_{k,0}[t])$  obtained by Monte Carlo simulation. Note that as  $r_{k,0}$  increases, also the delay  $\mathbb{E}[W_{k,0}]$  increases and the HARQ throughputs approach the genie-aided throughputs, in agreement with Theorem 3. The “o” marks indicate the throughput-delay points at which the HARQ protocol achieves 70%, 80% and 90% of the genie-aided throughput based on simulations. For example, under proportional fairness, 90% of the genie-aided throughput can be achieved at users  $(1, 0)$  and  $(18, 0)$  with average decoding delays of about 57 and 126 slots, respectively. These points (obtained by full system simulation) are accurately predicted by the analytical formulas of Section IV fitted with the Monte Carlo estimation of the probabilities  $\mathbb{P}(\mathcal{A}_{k,0}[t])$ .

For  $K = 36$  users per cell and  $M = 2$  BS antennas, assuming that exactly  $M = 2$  users are served in each slot, a round-robin scheduling with no outage (genie-aided rate allocation) would take an average delay of 18 slots. Remarkably, under proportional fairness, 90% of the genie-aided throughput can be achieved with about 57 slots of average delay for center user. This is only  $\approx 3$  times that of the genie-aided round-robin scheduling. For edge users, this is achieved with  $\approx 126$  slots of average delay for the edge users, which is only 7 times that of round-robin. Under max-min fairness, both users  $(1, 0)$  and  $(18, 0)$  achieve genie-aided throughputs close to 0.25 bits/channel use. The decoding delay for the center user is larger than for the edge user due to the fact that center users are scheduled very rarely. For the 70% point, edge users achieve 0.16 bits/channel use with average delay of 18 slots while center users achieves a similar throughput of 0.18 bits/channel use with delay of 44 slots.

## VI. CONCLUDING REMARKS

In this work we considered decentralized downlink scheduling in a multi-cell environment with multi-antenna BSs, where the scheduler at each BS has perfect CSIT about its own users and *statistical* information about the ICI caused by the other cells. Since each BS modifies its transmit covariance matrix at every slot, the ICI powers experienced at the users' receivers are random variable. We addressed the scheduling problem in the presence of uncertain ICI powers in the framework of stochastic network optimization. A straightforward application of this framework yields a conventional scheme based on adaptive variable-rate coding at the physical layer, and ARQ at the Logical Link Control layer. Then, a new combination of the same stochastic network optimization framework with incremental redundancy Hybrid ARQ at the physical layer was shown to improve over the conventional scheme, and achieve a network utility arbitrarily close to the performance of a genie-aided system that can schedule the user rates equal to the (non-causally known) instantaneous mutual information on each slot. For this scheme, we also showed that all Nash equilibria of the multi-cell decentralized scheduling game yield network utility values that can be uniformly upper and lower bounded by virtual systems corresponding to “extremal” ICI statistics, where the lower bound corresponds to the case of deterministic ICI powers equal to their mean values, and the upper bound corresponds to the case where all interfering BSs transmit to a single user at full power (rank 1 interfering covariance matrices). These bounds stay at a fixed gap that is independent of the cellular system configuration, i.e., of the channel path gain coefficients and operating SNR. The proposed incremental redundancy HARQ can be implemented in practice by using Raptor codes [40] at the physical layer, and need no protocol overhead to communicate slot-by-slot rate allocation as in adaptive variable-rate coding. Hence, the proposed HARQ scheme is both easier to implement and performs significantly better than the conventional variable-rate coding scheme. Also, we hasten to say that our approach applies directly to a variety of possible configurations, including different MU-MIMO precoding schemes and network MIMO schemes with clusters of coordinated cells [46]. In this paper we considered the case of linear ZFBF and no cell clustering for the sake of clarity of exposition. The approach can also be extended to the case of non-perfect CSIT, following [18]. Here we focused on perfect CSIT for its simplicity and in order to focus on the random nature of ICI as the fundamental source of uncertainty in a multi-cell environment.

## REFERENCES

- [1] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, pp. 70–77, Jul. 2000.
- [2] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: A high efficiency high data rate personal communication wireless system," in *Proc. IEEE Vehic. Tech. Conf., VTC-Spring*, May 2000.
- [3] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," *unpublished manuscript*.
- [4] D. Tse and S. Hanly, "Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. on Inform. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.
- [5] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.
- [6] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, 2000.
- [7] <http://www.3gpp.org/article/lte-advanced>.
- [8] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.
- [9] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.
- [10] S. A. Ramprashad and G. Caire, "Cellular vs. network MIMO: A comparison including the channel state information overhead," in *Proc. IEEE Intern. Symp. on Personal, Indoor and Mobile Radio Commun., PIMRC*, Tokyo, Japan, Sep. 2009.
- [11] G. Foschini, K. Karakayali, and R. A. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proc. Commun.*, vol. 152, no. 4, pp. 548–555, Aug. 2006.
- [12] F. Boccardi, F. Tosato, and G. Caire, "Precoding Schemes for the MIMO-GBC," in *Proc. Int. Zurich Seminar on Commun.*, Feb. 2006, pp. 10–13.
- [13] T. Marzetta, "How Much Training is Required for Multiuser MIMO ??" *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pp. 359–363, 2006.
- [14] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *submitted to IEEE Trans. on Inform. Theory*, Nov. 2007, Arxiv preprint cs.IT/0711.2642v2.
- [15] P. Ding, D. Love, and M. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. on Sig. Proc.*, vol. 55, no. 7, Part 1, pp. 3417–3428, Jul. 2007.
- [16] M. Kobayashi, G. Caire, and N. Jindal, "How much training and feedback are needed in MIMO broadcast channels??" in *Proc. IEEE Int. Symp. on Inform. Theory, ISIT*, Jul. 2008, pp. 2663–2667.
- [17] H. Shirani-Mehr and G. Caire, "Channel State Feedback Schemes for Multiuser MIMO-OFDM Downlink," *to appear in IEEE Trans. on Commun.*
- [18] H. Shirani-Mehr, G. Caire, and M. J. Neely, "Mimo downlink scheduling with non-perfect channel state knowledge," *Submitted to IEEE Transactions on Communications*.
- [19] L. Georgiadis, M. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*, ser. Foundations and Trends in Networking. Hanover, MA, USA: Now Publishers Inc., 2006, vol. 1, no. 1.
- [20] G. Fodor and C. Koutsimanis, "A low intercell interference variation scheduler for ofdma networks," in *Communications, 2008. ICC '08. IEEE International Conference on*, May 2008, pp. 3078–3084.

- [21] M. T. Ivrilac and J. A. Nossek, "Intercell-interference in the gaussian miso broadcast channel," in *GLOBECOM*. IEEE, 2007, pp. 3195–3199.
- [22] M. J. Neely, E. Modiano, and C. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE INFOCOM Proceedings*, March 2005.
- [23] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad-Hoc Networks*, vol. 23, no. 1, pp. 89 – 103, Jan 2005.
- [24] Q. Bi and S. Vitebsky, "Performance analysis of 3G-1X EVDO high data rate system," in *Proc. IEEE Wireless Commun. and Networking Conf., WCNC*, vol. 1, Mar. 2002, pp. 389–395.
- [25] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with Chase combining and incremental redundancy for HSDPA," in *Proc. IEEE Vehic. Tech. Conf., VTC-Fall*, vol. 3, Sep. 2001, pp. 1829–1833.
- [26] R. Love, A. Ghosh, W. Xiao, and R. Ratasuk, "Performance of 3GPP high speed downlink packet access (HSDPA)," in *Proc. IEEE Vehic. Tech. Conf., VTC-Fall*, vol. 5, Sep. 2004, pp. 3359–3363.
- [27] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. on Inform. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.
- [28] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.
- [29] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.
- [30] M. J. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2915–2934, 2006.
- [31] J. Jose, A. Ashikhmin, P. Whiting, and S. Vishwanath, "Scheduling and pre-conditioning in multi-user MIMO TDD systems," *Arxiv preprint cs.IT/0709.4513*.
- [32] P. Ding, D. Love, and M. Zoltowski, "Multiple Antenna Broadcast Channels With Shape Feedback and Limited Feedback," *IEEE Trans. on Sig. Proc.*, vol. 55, pp. 3417–3428, 2007.
- [33] J. W. Friedman, "A non-cooperative equilibrium for supergames," *Review of Economic Studies*, vol. 38, no. 113, pp. 1–12, 1971. [Online]. Available: <http://ideas.repec.org/a/bla/restud/v38y1971i113p1-12.html>
- [34] ——, *Oligopoly and the Theory of Games*. Amsterdam [u.a.]: North-Holland, 1977.
- [35] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965. [Online]. Available: <http://dx.doi.org/10.2307/1911749>
- [36] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *Vehicular Technology, IEEE Transactions on*, vol. 43, no. 2, pp. 359–378, May 1994.
- [37] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and simple new algorithm," *IEEE Trans. on Sig. Proc.*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.
- [38] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 3, pp. 528–541, March 2006.
- [39] H. Huh, H. Papadopoulos, and G. Caire, "MIMO broadcast channel optimization under general linear constraints," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 28 2009–July 3 2009, pp. 2664–2668.
- [40] A. Shokrollahi, "Raptor Codes," *IEEE Trans. on Inform. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.
- [41] J. Forney, G., "Exponential error bounds for erasure, list, and decision feedback schemes," *Information Theory, IEEE Transactions on*, vol. 14, no. 2, pp. 206–220, Mar 1968.

- [42] M. Sharif and B. Hassibi, "On the capacity of a mimo broadcast channel with partial side information," *IEEE Trans. on Inform. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.
- [43] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Downlink Channels with Limited Feedback and User Selection," *IEEE J. Select. Areas Commun.*, vol. 25, pp. 1478–1491, 2007.
- [44] G. Grimmet and D. Stirzaker, *Probability and Random Processes*. Oxford Univ. Press, 2004.
- [45] M. Abramowitz and I. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Dover Publications, 1965.
- [46] G. Caire, S. A. Ramprashad, H. C. Papadopoulos, C. Pepin, and C.-E. W. Sundberg, "Multiuser MIMO downlink with limited inter-cell cooperation: Approximate interference alignment in time, frequency and space," *Proc. of Forty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2008.

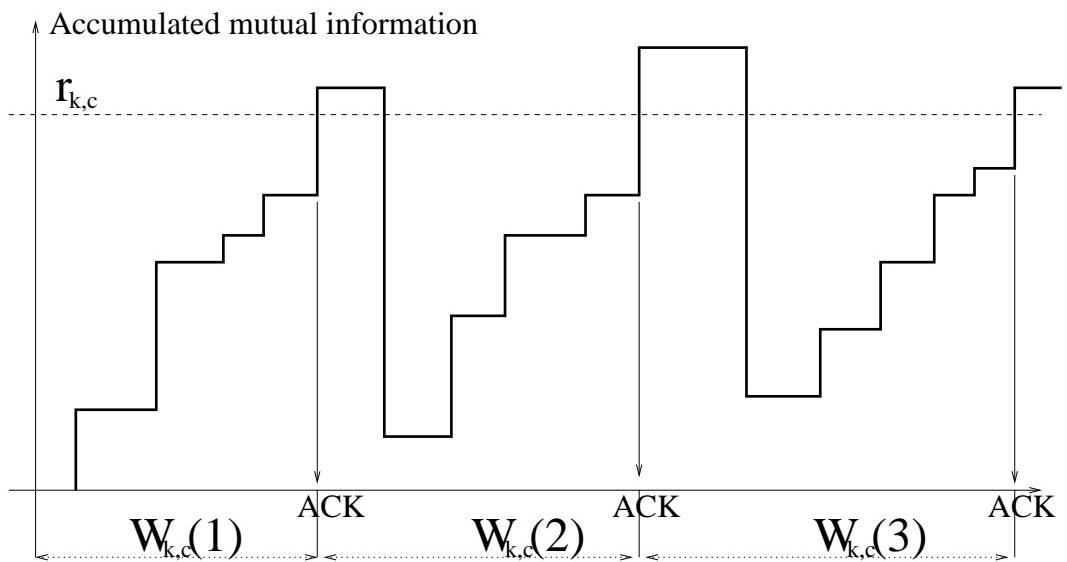


Fig. 1. Qualitative plot of the mutual information level-crossing process that determines the decoding events of the HARQ protocol. The jumps of the accumulated mutual information process correspond to slot times at which user  $(k, c)$  is active.

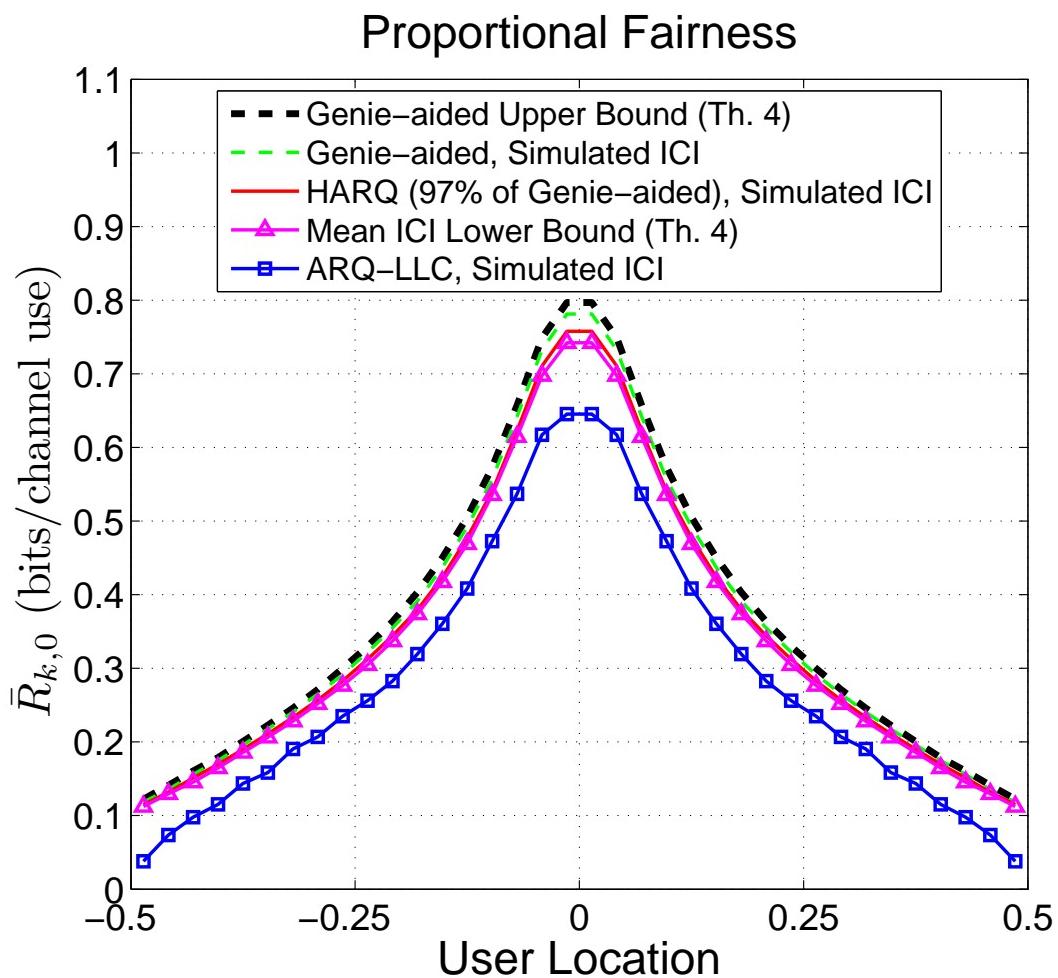


Fig. 2. Average-throughput, proportional fairness.

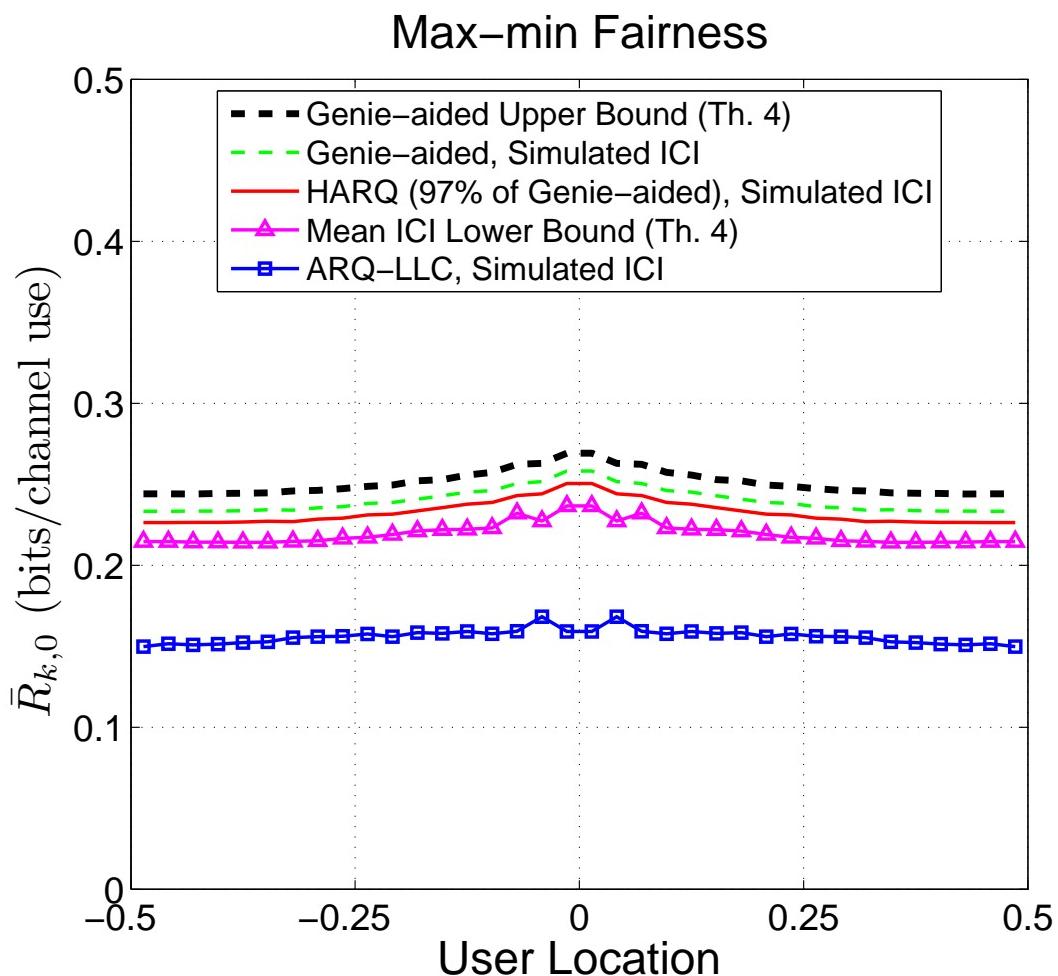


Fig. 3. Average-throughput, max-min fairness.

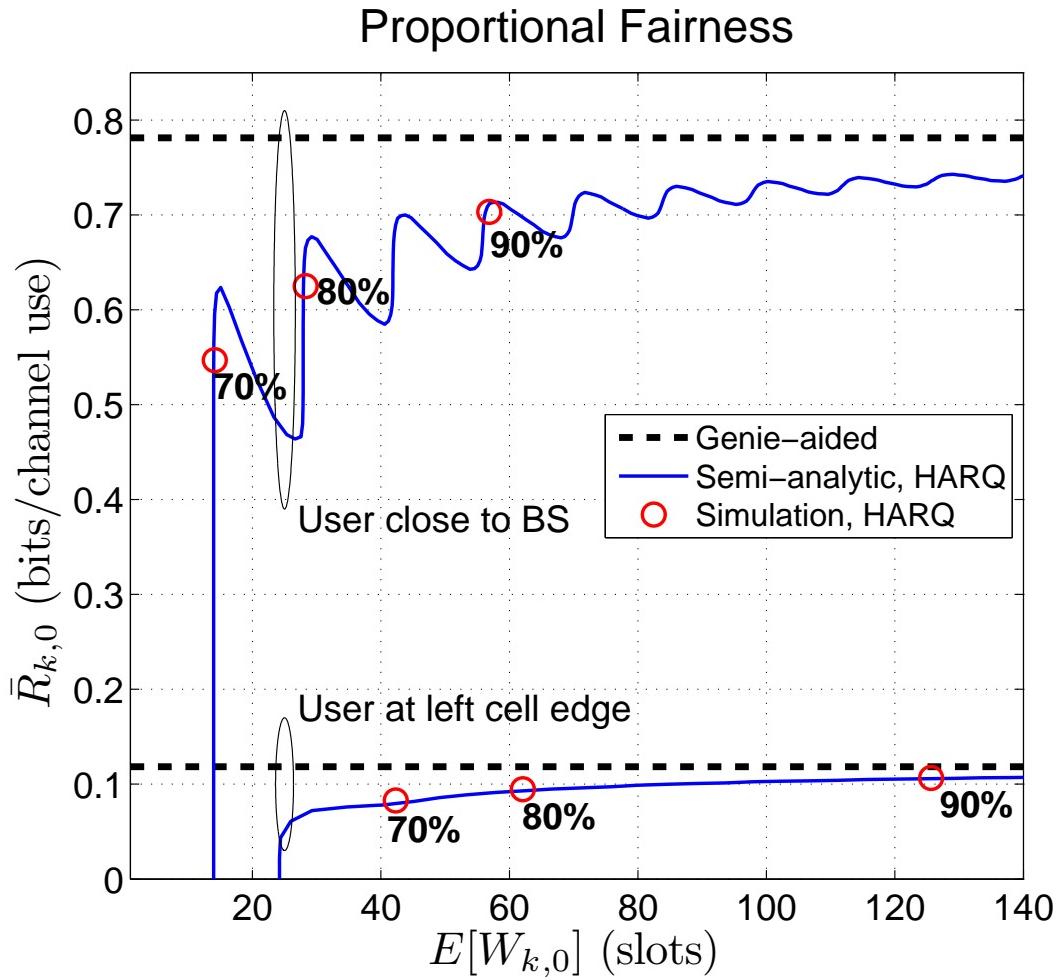


Fig. 4. Average rate vs. decoding delay with proportional fairness for two sample users.

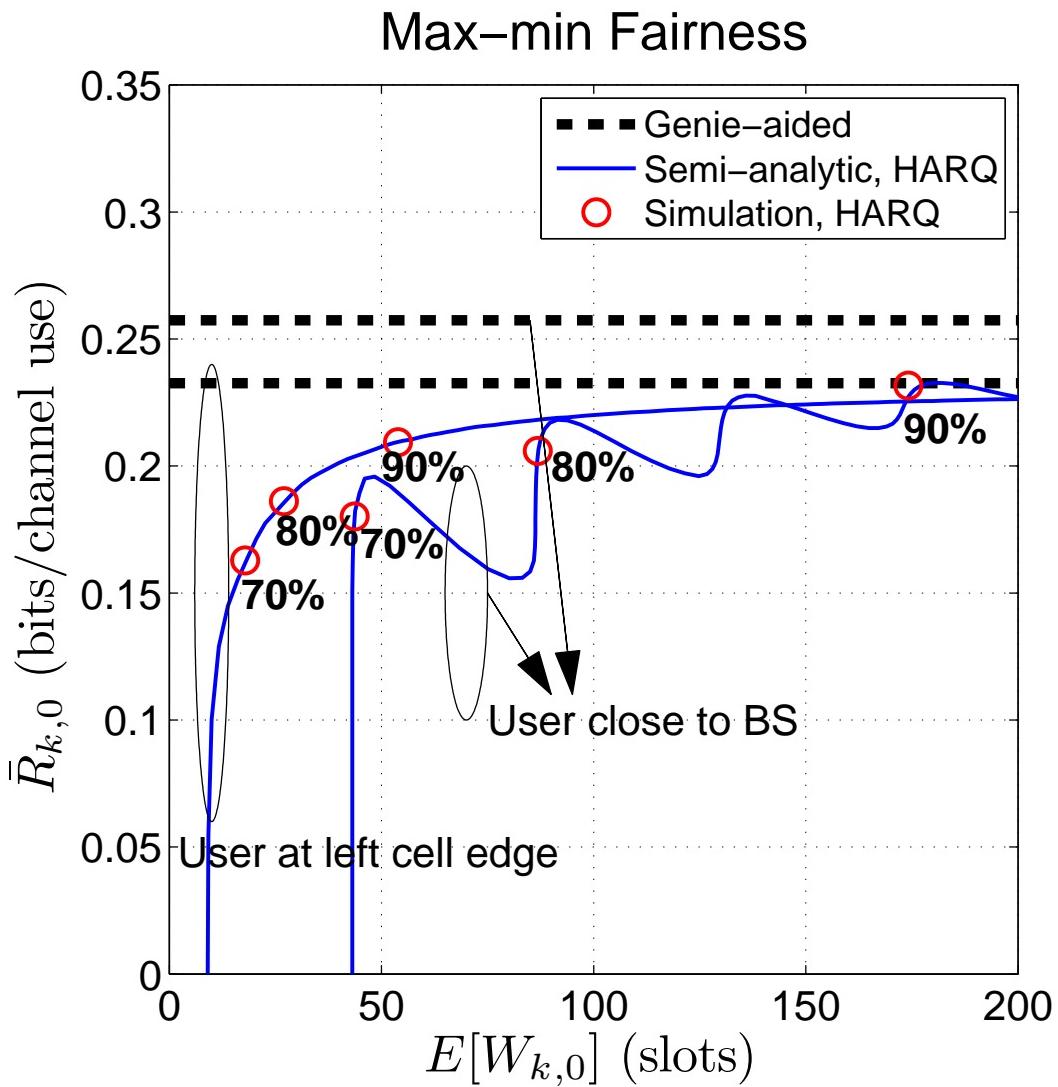


Fig. 5. Average rate vs. decoding delay with max-min fairness for two sample users.